de Sousa, F., Foster, P. G., Donoghue, P., Schneider, H., & Cox, C. J. (2019). Nuclear protein phylogenies support the monophyly of the three bryophyte groups (Bryophyta Schimp.). *New Phytologist*, *222*(1), 565-575. https://doi.org/10.1111/nph.15587

Nuclear protein phylogenies support the monophyly of the three bryophyte groups (Bryophyta Schimp.)

Filipe de Sousa[1], Peter G. Foster[2], Philip C. J. Donoghue[3], Harald Schneider[4,2,3], Cymon J. Cox[1*]

[1]Centro de Ciências do Mar, Universidade do Algarve, Gambelas, 8005-319 Faro, Portugal
[2]Department of Life Sciences, Natural History Museum, London SW7 5BD, United Kingdom
[3]School of Earth Sciences, University of Bristol, Bristol BS8 1TQ, United Kingdom
[4]Center of Integrative Conservation, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Yunnan 666303, China
*Corresponding author email: cymon.cox@googlemail.com Tel.: (+351)289800051 ext 7380

total word count: 6177
abstract: 200
introduction: 1432
material and methods: 1557
results: 864
discussion: 2065
acknowledgements: 59

number of figures: 2
figures in colour: Fig. 1, Fig. 2
number of supplemental files: 1

ORCID of authors: Filipe de Sousa: 0000-0003-4681-8951; Peter G. Foster: 0000-0003-0194-9237; Philip C.J. Donoghue: 0000-0003-3116-7463; Harald Schneider: 0000-0002-4548-7268; Cymon J. Cox: 0000-0002-4927-979X

acceptance date: 31 October 2018

CCMAR twitter: https://twitter.com/CienciasDoMar

Summary

- Unraveling the phylogenetic relationships between the four major lineages of terrestrial plants (mosses, liverworts, hornworts, and vascular plants) is essential for an understanding of the evolution of traits specific to land plants, such as their complex life cycles, and the evolutionary development of stomata and vascular tissue.
- Well supported phylogenetic hypotheses resulting from different data and methods are often incongruent due to processes of nucleotide evolution which are difficult to model: for example, substitutional saturation and composition heterogeneity. We reanalyse a large published dataset of nuclear data and model these processes using degenerate codon recoding and tree-heterogeneous composition substitution models.
- Our analyses resolve bryophytes as a monophyletic group and show that the non-monophyly of the clade, that is supported by the analysis of nuclear nucleotide data, is due solely to fast-evolving synonymous substitutions.
- The current congruence among phylogenies of both nuclear and chloroplast analyses lend considerable support to the conclusion that the bryophytes are a monophyletic group. An initial split between bryophytes and vascular plants implies that the bryophyte life cycle (with a dominant gametophyte nurturing an unbranched sporophyte) may not be ancestral to all land plants and that stomata are likely a symplesiomorphy among embryophytes.

Keywords:
bryophytes; compositional heterogeneity ; phylogenomics ; substitutional saturation ; life cycle; land plants

Introduction

Plants are the main primary producers in terrestrial environments, constituting the majority of above-ground biomass and representing a major atmospheric carbon-sink that has shaped the climate globally (Lenton *et al.* 2012). However, despite their ecological importance for life on land, the evolutionary relationships of the major lineages of terrestrial plants and their immediate ancestors is not yet fully understood. In particular, the relationships among the three bryophyte groups, namely mosses, liverworts, and hornworts, and their relationship to the vascular plants (tracheophytes) have long been controversial (reviewed by Cox, 2018). Land plants develop via a sporophytic embryo that is nurtured by the gametophyte, and hence are collectively referred to as embryophytes. The freshwater charophyte green algae have for a long time been recognized as the closest living relatives of the embryophytes (Karol *et al.* 2001; McCourt *et al.* 2004) and recent molecular evidence suggests that the Zygnematales (Timme *et al.* 2012; Civán *et al.* 2014) or a clade including the Zygnematales and Coleochaetales (Wodniok *et al.* 2011; Laurin-Lemay 2012) share the most recent common ancestor with the embryophytes.

The evolution of land plants was accompanied by a shift from a haplobiontic lifecycle with a single multicellular haploid gametophytic generation, as seen today in freshwater charophytes, to a diplobiontic lifecycle, characterized by an alternation of multicellular haploid and diploid generations (Niklas & Kutschera 2010). In all extant land plants, embryonic sporophytes are dependent on parental gametophytic tissue for at least part of their development (Graham and Wilcox 2000), but two contrasting diplobiontic life strategies can be distinguished: in bryophytes, the haploid gametophytes are the dominant vegetative stage, whereas in tracheophytes (lycophytes, ferns, and seed plants), the diploid sporophyte is the main vegetative stage (Niklas & Kutschera 2010). In the absence of a well-supported phylogenetic hypothesis on the relationships and order of divergence of early land plants, is it not possible to determine which type of lifecycle characterized their common ancestor. If tracheophytes are derived from a bryophyte ancestor, the ancestral lifecycle of embryophytes would probably have been predominantly gametophytic (Niklas & Kutschera 2010 ; Ligrone *et al.* 2012). If, instead, the first split occurred between bryophytes and tracheophytes, then the embryophyte ancestor could have had diplobiontic lifecycle (Stebbins and Hill 1980), with stomata possibly arising in the ancestral sporophyte of all land plants.

The transition of ancestral plants to land, from an aquatic environment, is thought to have occurred c. 480 Ma, in the late Silurian period (Kenrick *et al.* 2012; Magallón *et al.* 2013) but recent estimates have dated this transition earlier to 515.1- 470.0 Ma in the late Cambrian or early Ordovician (Morris *et al.* 2018). However, without a reliable phylogenetic hypothesis, an accurate

dating of the origin of the embryophytes is more difficult to establish (Morris *et al.* 2018). To date, the most widely accepted evolutionary hypothesis is that the tracheophytes derive from an early bryophyte lineage, and that either liverworts alone (Karol *et al.* 2001; Qiu *et al.* 2006; Gao *et al.* 2010; Karol *et al.* 2010; Clarke *et al.* 2011;), liverworts plus mosses (Karol *et al.* 2010), or the hornworts alone (Nishiyama and Kato 1999; Wickett *et al.* 2014), are the sister-group to the remaining land plants. However, the view that bryophytes form a monophyletic group, while receiving less frequent acceptance, has not been ruled out (Nishiyama *et al.* 2004; Cox *et al.* 2014; Wickett *et al.* 2014; Morris *et al.* 2018; Puttick *et al.* 2018; Nishiyama *et al.* 2018).

The absence of a definitive phylogeny of land plants, in spite of the considerable amount of data available from all three genomic compartments, is due to the challenges posed when comparing anciently diverged molecular data. Regardless of the origin of the data, two main factors are known to cause systematic error in phylogenetic reconstruction of ancient phylogenies: high substitution rates (ultimately leading to substitution saturation and loss of phylogenetic signal) and composition biases among sites and between taxa (data- and tree-heterogeneity, respectively; Liu *et al.* 2014). Substitutional saturation occurs when multiple substitutions at the same site overwrite synapomorphies and create homoplasies (Philippe *et al.* 2011) thereby generating “noisy” data that can affect branch support and lead to erroneous phylogenetic inference (Jeffroy & Brinkmann 2006). Saturation is dependent on time and substitution rate, and is therefore more pronounced in faster-evolving nucleotide data (Liu *et al.* 2014). Methodological approaches for alleviating the problem of substitutional saturation include removing third codon positions (Wickett *et al.* 2014), which corresponds in most cases to the removal of fast-evolving synonymous substitutions, and using codon degeneracy, which effectively removes all synonymous substitutions by recoding synonymous nucleotides at codon sites with nucleotide ambiguity codes (Cox *et al.* 2014).

Nucleotide or amino acid compositions are generally modeled as their respective frequencies at equilibrium, and include the probability of change from one state to another. The Markov models used as substitution models in phylogenetics assume a stationary process that does not vary across time or across the data. However, we often see that different genes (or data partitions) have different compositions, which violates the assumption that the process does not differ over the data. We can relax this assumption and model composition heterogeneity among data by applying different Markov models, with different compositions, to different data partitions. Furthermore, compositional heterogeneity among taxa is also often seen at all levels of phylogenetic organisation, in violation of the assumption that the process does not vary across the tree (or over time). Such heterogeneity may be caused by differences in direct selective pressures or by variation in passive mutation processes. We can sometimes ameliorate this heterogeneity by judicious site- or

taxon-stripping, or alternatively we can accommodate the heterogeneity by using appropriate tree-heterogeneous composition substitution models (Foster 2004; Inagaki *et al.* 2004; Inagaki and Roger 2006; Regier *et al.* 2010; Rota-Stabelli *et al.* 2012; Lockhart *et al.* 1992; Mooers and Holmes 2000; Blanquart and Lartillot 2008). Indeed, homogeneity of the substitution process should always be verified in molecular data used to reconstructing ancient phylogenies, and, if the data are shown to be non-stationary, then appropriate tree-heterogeneous composition substitution models should be used (Foster *et al.* 2009; Liu *et al.* 2014; Cox *et al.* 2014). If stationary substitution models are applied to composition tree-heterogeneous data, an artificial, but possibly statistically well-supported, clustering of taxa with similar compositions may occur (e.g. Foster 2004; Cox *et al.* 2008). Moreover, differences in composition at the nucleotide level are reflected at codon level in the form of different synonymous codon preferences among lineages, or codon-usage bias (Gouy and Gautier 1982; Stenøien 2005; Inagaki *et al.* 2004; Inagaki and Roger 2006; Zhou and Li 2009; Plotkin and Kudla 2011; Rota-Stabelli *et al.* 2012; Liu *et al.* 2014) which may strongly impact phylogenetic reconstruction when using codon models if shared codon preference is mistaken for shared ancestry (Inagaki *et al.* 2004; Inagaki and Roger 2006; Regier *et al.* 2010; Rota-Stabelli *et al.* 2012; Cox *et al.* 2014). Differences in codon-usage occur between species but also within genomes, and can be a consequence of translational selection, as well as differences in mutational bias (Bulmer 1988; Sharp *et al.* 1993). A possible approach to mitigate the effect of amino acid composition bias on phylogenetic reconstruction is to re-code protein data by defining amino acid groups that show similar substitution properties (Susko and Roger 2007; Rota-Stabelli *et al.* 2012).

In this study we analyse molecular sequence data from the nuclear genome to clarify relationships among land plant lineages using novel analytical approaches. We assume the monophyly of tracheophytes and of each of the three bryophyte lineages, which has been consistently demonstrated (Qiu *et al.* 2006; Chang and Graham 2011; Liu *et al.* 2014; Wickett *et al.* 2014). We attempt to balance representatives of each bryophyte and tracheophyte lineage, to achieve greater tree symmetry, as asymmetrical trees are less likely to be correctly estimated than symmetrical trees, due to the shorter average branch length which expands the number of anomalous gene trees (Huang and Knowles 2009). More balanced sampling among lineages is also likely to minimise the effect of long-branch attraction, which often influences deep phylogenetic relationships (Phillipe and Laurent 1998). We revisit a large published dataset of nuclear loci (Wickett *et al.* 2014) and implement complete degenerate recoding of synonymous substitutions to the whole data set. To be able to apply complex and computationally challenging substitution models we also constructed a smaller data set with selected loci (100) and a reduced number of taxa (26). We test these data using heterogeneous models of substitution that accommodate mutational

heterogeneity and show that analyses using the best-fitting composition models support the monophyly of bryophytes.

Materials and Methods

Analyses of Wickett et al. data (620 genes, 103 taxa)

The data of Wickett *et al.* (2014), consisting of 620 nuclear genes and 103 taxa was obtained from a public data repository (Wickett *et al.* 2015. Onekp_pilot. Retrieved from http://www.cyverse.org). The original data matrix (labeled FNA2AA.trim50genes50sites.allPos.unpartitioned.phylip) consisted of 436,077 sites of in-frame coding sequence, after genes missing more than 50% of taxa and sites with more than 50% of gaps were removed. Synonymous versus non-synonymous substitution rates of the 85 of 620 genes that were "gapless" were calculated in PAML (vers. 4.6; Yang, 2007). The concatenated 620 gene data set was recoded with codon-degenerate characters using the script (recode_matrix.py; Li pers. comm.), which places ambiguity characters at synonymous third codon positions, at first codon positions of amino acids Leucine (L) and Arginine (R), and at both first and second codon positions of amino acid Serine (S), which can be coded with either purines (AG) or pyrimidines (TC) at these positions. All third codon positions were removed from both the original and the recoded matrices (290,718 sites). The aminoacid translation matrix (labeled FAA.trim50genes50sites.clustered.partitioned.phylip) was also obtained. Hence there were three derived data matrices based on the original taxon and gene selection of Wickett et al: 1) original data matrix without third codon positions, 2) original data with codon-degenerate recoding and without third codon positions, and 3) the amino acid translation of the original matrix.

Maximum likelihood bootstrap analyses were conducted on all matrices using RAxML (MPI-compiled vers. 8.2.8; Stamatakis 2014) using the "full" (RAxML notation: -b) bootstrap algorithm and 200 replicates. The original nucleotide data matrix (436,077 sites) was analysed by bootstrapping with a general time-reversible model of substitution (GTR), with a discrete (4 categories) gamma distribution of among-site rate variation (G_4) with empirical composition values (F_{emp}) with 200 bootstrap replicates (RAxML notation: GTRGAMMA). The data sets without third codon positions (290,718 sites), and the same matrix but with codon-degenerate coding, were analysed by bootstrapping with a GTR+G_4 with the composition estimated via ML(F_{est}) (RAxML notation: GTRGAMMAX). The latter data set (no third codon positions, codon-degenerate coding)

was also analysed using a GTR model but with the Per Site Rate model (PSR; Stamatakis and Aberer 2013) (previously named the CAT-rates approximation), each with ML estimated composition frequencies (F_{est})(RAxML notation: GTRCATX). Analyses of the original and derived matrices were conducted to compare the effect of third codon position removal with the effect of synonymous substitutions, the latter through the use of codon-degenerate recoding that effectively eliminates synonymous substitutions at first and second codon positions. For the concatenated gene protein translation data (145359 sites), the partitioning scheme calculated by Wickett *et al.* (2014) (9 categories; file labeled: "PARTITION_FOR_W14_AA_103t_145359aa.partition") was used (RAxML notation: -q) with both the G_4 and PSR rate category estimations and F_{est} (RAxML notation: PROTGAMMA<>X and PROTCAT<>X, where <> is an arbitrary model that is ignored) and 100 bootstrap replicates.

Gene and taxon selection for the reduced data set (100 genes, 26 taxa)

Using non-stationary substitution models for phylogenetic inference requires substantial computational capacity, and it was therefore necessary to reduce the sampling of genes and taxa. We chose to select the genes that had the lowest composition heterogeneity among taxa and the shortest tree lengths, to minimize composition effects and substitutional saturation. Out of the 620 genes in the original nucleotide matrix, we analysed those larger than 500 bp (388 genes), in MrBayes (vers. 3.2.6; Ronquist *et al.* 2012), under the composition homogeneous GTR+G_4 model of nucleotide substitution. Markov-Chain Monte Carlo (MCMC) analyses were run for 500,000 generations, after which a stop-rule was employed with the default 0.05 for the average standard deviation of split frequencies (ASDOS). Out of 388 genes, 43 did not converge (ASDOS < 0.05). Composition homogeneity tests of posterior predictive distributions of the chi-square (X^2) statistic were conducted using p4 (vers. 1.2.0; Foster 2004) indicated that all 345 genes were significantly non-homogeneous ($p<0.05$). Genes were scored for their X^2 value of composition homogeneity and for mean tree lengths of sampled trees from the posterior tree distribution, and ranked by both scores. The mean of ranks was used as a final ranking, and the 100 genes with the lowest chi-square and tree lengths were selected.

Taxa were scored in the selected 100 genes for number of genes in which they were present and for the total percentage of missing sites. For each taxon, the absolute %GC deviation from the mean of entire gene alignment composition was also calculated. These values were used, in each of the six main land plant groups, and in the outgroups, to select the most appropriate taxa in order to minimise both %GC deviation and number of missing taxa, resulting in a final list of 26

taxa. The concatenated 100 gene and 26 taxa nucleotide alignment comprised 69,903 sites and the translated amino acid alignment, obtained with the alignment program SeaView (vers. 4.5.4; Gouy *et al.* 2009), comprised 23,301 sites.. A matrix with complete codon-degeneracy was obtained from the concatenated nucleotide alignment. The concatenated amino acid matrix was recoded into Dayhoff amino acid groups (6 groups: c, stpag, ndeq, hrk, milv, fyw; Dayhoff *et al.* 1978) using the program P4. Individual nucleotide and amino acid matrices of the 100 genes were also generated.

Phylogenetic analyses of the reduced data set (100 genes, 26 taxa)

To assess the effect of synonymous substitutions, both the concatenated nucleotide and the codon-degenerate data matrices of the 100 gene and 26 taxa reduced data set were analysed under the GTR+G_4+F_{est} model of substitution (RAxML notation: GTRGAMMAX), with 300 bootstrap replicates, in RAxML. The nucleotide data alignment was also analysed in PhyloBayes MPI (vers. 1.6; Lartillot *et al.* 2009) using the model CAT-GTR+G_4 to assess the effect of among-site composition heterogeneity. To test the effect of data partitioning under maximum-likelihood, genes were grouped into partitions using the "greedy" algorithm in IQ-TREE (multicore vers. 1.5.3; Nguyen *et al.* 2015; Chernomor *et al.* 2016). A bootstrap analysis with 100 replicates of the 9 optimal partitions was performed using IQ-TREE (see Supporting Information, Fig. S7 for details). We then tested whether the phylogenetic signal obtained from the analyses of nucleotide data differed from the signal obtained from the analyses that use models and data transformations aimed at mitigating the effect of homoplasy due to saturation. These analyses were performed: 1) on nucleotide data under codon models; 2) on amino acid matrices; 3) on matrices of grouped amino acids. Codon analyses were performed on the 100 gene dataset using IQ-TREE, with 100 bootstrap replicates using the models GY2K+F3X4+G_4 and MG2K+F3X4+G_4. An optimal model for the concatenated amino acid data set was determined using Modelgenerator (vers. 0.85; Keane *et al.* 2006). Bootstrap analysis were performed in RAxML under the LG+G_4+F_{est} (RAxML notation: PROTGAMMALGX) model, with 300 replicates on both the amino acid and Dayhoff-recoded data sets. The amino acid dataset was also analysed in PhyloBayes under the CAT-LG+G_4 model with 2 parallel MCMC runs.

Bayesian MCMC analyses of individual nucleotide and amino acid data matrices of the reduced 100 genes, 26 taxon set were performed using P4. Nucleotide data were analysed under the GTR+G_4 model of substitution. Models for analysing individual amino acid matrices were inferred in Modelgenerator. Each matrix was analysed assuming both composition homogeneity (F_{CV1}: one composition vector) and heterogeneity ($F_{CV>1}$: two or more composition vectors) using the node-

discrete composition heterogeneity model (NDCH; Foster 2004; Cox *et al.* 2008) which accounts for base composition differences between branches on a tree.

To assess the effect of composition heterogeneity we analysed the concatenated nucleotide, amino acid, and Dayhoff group matrices with Bayesian MCMC using both tree-homogeneous and tree-heterogeneous composition models. The concatenated and codon-degenerate nucleotide matrices of the 100 gene, 26 taxon set were analysed with Bayesian MCMC using the composition homogeneous model $GTR+G_4+F_{CV1}$ and composition heterogeneous NDCH model ($GTR+G_4+F_{CV>1}$) in P4. The concatenated amino acid alignment was analysed using the composition homogeneous model $LG+G_4+F_{CV1}$, and the Dayhoff-recoded data were analysed under the $GTR+G_4+F_{CV1}$ model. Composition heterogeneous NDCH model analyses were conducted on the concatenated amino acid data ($LG+G_4+F_{CV>1}$) and the Dayhoff-recoded data set ($GTR+G_4+F_{CV>1}$). A minimum of two runs were performed for each analysis. Run convergence was assessed by estimating ASDOS, which was accepted when lower than 0.05, by plotting the MCMC sample likelihoods, and comparing marginal likelihoods. Effective sample size (ESS) values and acceptances for proposals were estimated and assessed using P4 methods. The fit of the composition models was determined during the MCMC by posterior predictive simulations of the X^2 statistic of composition homogeneity (Foster 2004). Marginal likelihoods were estimated in P4 following the Eqn 16 method of Newton and Raftery (1994). Bayes factors, which are used to compare the relative adequacy of competing models (Nylander *et al.* 2004), were estimated from the log-marginal likelihood of analyses using homogeneous (null) and non-homogeneous (alternative) models, when the alternative model was accepted under posterior predictive simulation. Alternative models that had a high log-Bayes Factors ($\log_e$ BF>10 units), calculated as $2*(\log_e L(\text{alternative model}) - \log_e L(\text{null model}))$ were considered better-fitting than the homogeneous model. A PhyloBayes analysis using the $CAT\text{-}LG+G_4$ model was conducted on the concatenated amino acid data.

Analyses were performed on the CCMAR computational cluster facility GYRA at the University of Algarve or INGRID part of the Infraestrutura Nacional de Computação Distribuída (INCD) in Portugal. Details of each analysis are presented in the legends of Figures S1–S17 in the Supporting Information.

Results

Wickett et al. nucleotide and amino acid data analyses

The analysis of the 620 gene nucleotide dataset using maximum likelihood resulted in a tree that supports hornworts as the sister-group to the remaining land plants with a bootstrap support (BS) of 89% (Fig. S1). The same supported relationship (BS=98%) is shown when nucleotides at third codon positions are excluded from the data (Fig. S2). This result is concordant with the equivalent analysis of the 620 genes dataset in Wickett *et al.* (2014; their Fig. 2) with third codon positions excluded.

Analysing the 620 gene dataset with codon-degenerate recoded data and excluded third codon positions, using maximum likelihood and the GTR+G_4 model, resulted in trees showing bryophytes as a monophyletic group, albeit with low support (BS=54%, Fig. S3). Using the GTR+PSR rate model, however, yields a tree that supports the paraphyly of bryophytes (BS=85%) and showing hornworts as the sister-group to all other land plants (Fig. S4). Similarly, differences between rate models were also observed in the maximum likelihood bootstrap analyses of partitioned amino acid data, which identifies hornworts as the sister-group to all other embryophytes when the PSR rate model is used (BS=75%, Fig. S5) but resolves the three bryophyte lineages as a monophyletic group when the G_4 rate model is used (BS=76%, Fig. S6).

Reduced nucleotide data set analyses (100 genes, 26 taxa)

None of the 100 individual protein-coding genes (>500bp) analysed had a stationary homogeneous composition across the tree. Most genes had a best-fitting model with two composition vectors (F_{CV2}), and five genes were better fitted by three vectors (F_{CV3}). Of the 100 individual amino acid gene translations analysed, 24 were compositionally tree-homogeneous, while the remaining protein models required up to six composition vectors (F_{CV6}) to fit the data (Table S1).

All maximum likelihood analyses of the reduced nucleotide dataset (100 genes, 26 taxa) show full support (BS=100%) for the monophyly of embryophytes and of each of its four major lineages (mosses, liverworts, hornworts, vascular plants). When the data are analyzed using the GTR+G_4 model the resulting tree supports hornworts as sister-group to the remaining embryophytes (BS=81%; Fig. 1a). Analysis of the partitioned data using IQTREE also places hornworts as the sister-group to the remaining embryophytes with low bootstrap support (BS=68%; Fig. S7). In contrast, when the data were analyzed using degenerate coding for all synonymous codon positions, the resulting tree showed the three bryophyte lineages forming a well-supported monophyletic group (BS=89%; Fig. 1b).

Bayesian analyses of the reduced nucleotide dataset using both tree-homogeneous (F_{CV1}) and tree-heterogeneous NDCH (F_{CV2}) composition models show hornworts strongly

supported as the sister-group to the remaining land plants (PP=1.0; Figs. S8 and S9, respectively). Although the two runs of the heterogeneous analysis did not converge, they both recovered the same topology (Fig. S9): here we report only the diagnostic values of the MCMC with the highest likelihood. The model with two composition vectors (F_{CV2}) fit the data with a posterior predictive simulation X^2 distribution of the composition homogeneity statistic (p=1.0), whereas the homogeneous (F_{CV1}) model was rejected (p=0.0). The Bayes factor comparing the composition homogeneous and heterogeneous models strongly support the heterogeneous model ($2\log_e$ BF=9016.7). Bayesian reconstructions using the PhyloBayes CAT model resulted in a tree showing mosses as the sister-group to other land plants (PP=0.99; Fig. S10), which contrasts with all other results obtained from the same data. Analyses of the degenerate-recoded data with both a homogeneous (F_{CV1}) and heterogeneous model (F_{CV2}) show bryophytes as a monophyletic group with maximum support (PP=1.0; Figs. S11 and S12, respectively). Posterior predictive simulations of composition fit to the data reject the homogeneous model (p=0.0) but not heterogeneous model (p=0.99). The Bayes factor strongly favours the heterogeneous model ($2\log_e$ BF=961.3). Maximum likelihood bootstrap analyses of the codon-site data using models GY2K and MG2K place hornworts as the sister-group to other land plants with full bootstrap support (BS=100%; Figs. S13 and S14, respectively).

Reduced amino acid data analyses (100 genes, 26 taxa)

Maximum likelihood bootstrap analysis of the amino acid dataset using the $LG+G_4$ model resulted in a tree showing monophyletic bryophytes but with low bootstrap support (BS=56%; Fig. 2a). However, a similar analysis with the data recoded into Dayhoff groups resulted in higher bootstrap support for a monophyletic bryophyte clade (BS=80%, Fig. 2b). Bayesian MCMC analyses of the concatenated amino acid dataset using both tree-homogeneous and NDCH tree-heterogeneous models recovered the bryophytes as monophyletic (Figs. 2c and 2d, respectively). However, whereas the poor-fitting (X^2=0.0) homogeneous model showed low support (PP=0.84; Fig. 2c), the best-fitting NDCH composition model (F_{CV5}) had a highly significant posterior probability for monophyletic bryophytes (PP=0.98; Fig. 2d). The Bayes factors comparing the tree-homogeneous and tree-heterogeneous composition models strongly favour the latter model ($2\log_e$ BF=1513.9). Bayesian MCMC of the Dayhoff-recoded dataset resolve bryophytes as monophyletic with full branch support under both homogeneous and heterogeneous models (PP=1.0, Figs. S15 and S16, respectively). Posterior predictive simulations of the composition reject the homogeneous model (F_{CV1}; P=0.0) and support a model with two composition vectors (F_{CV2};

P=0.99), and the Bayes factor strongly favours the heterogeneous model ($2\log_e$ BF=400.5). A PhyloBayes analysis of the amino acid data using the model CAT-LG+G_4 also yields a tree that supports the monophyly of bryophytes (PP=0.99; Fig. S17).

Discussion

The effect of degenerate-codon re-coding on fast-evolving nucleotide data

The recoding of nucleotide alignments with codon-degenerate ambiguity codes negates the effect of not only synonymous substitutions at third codon positions, but also those at second- and first-codon positions in L, R, and S codons, while still retaining those non-synonymous substitutions that are eliminated by the common practice of deleting third codon positions. Synonymous substitutions experience less selection than nonsynonymous substitutions and have previously been shown to range between 2-40X faster than non-synonymous substitutions in nuclear genes (Yang and Neilsen 1998). In both the 620 and 100 gene datasets analysed here, synonymous substitutions occurred a mean of ~12.5X (ω=dn/ds=~0.08) faster than non-synonymous substitutions, ranging between 3-400X (ω=dn/ds=0.3492 - 0.0025) and 6-300X (ω=dn/ds=0.1742 - 0.0033) faster in the 620 and 100 gene datasets respectively (see Supporting Information Notes S1 and S2). Homologous sites among taxa at which synonymous substitutions occur are therefore more likely to exhibit substitution saturation and hence character homoplasy across the phylogeny, which is compounded by convergent compositional biases due to different mutation pressures among taxa (Cox *et al.* 2014).

Codon-degenerate recoded nucleotide data resulted in inferred topologies that differ from those obtained from complete alignments and from alignments with all third codon positions removed. Simply excluding third codon positions from the 620 gene dataset recovers hornworts as the sister-group to the remaining embryophytes (Fig. S2), as reported by Wickett *et al.* (2014, their Fig. 2). However, when the L, R, and S synonymous codons (which include synonymous substitutions at first- and second-codon positions) are recoded with ambiguity codes (ie codon-degenerate recoding), in addition to the exclusion of third codon positions, the resulting tree shows bryophytes as monophyletic (Fig. S3). This results indicates that although most saturated sites occur at third codon positions, the effect of synonymous substitutions at first- and second-codon positions in L, R, and S amino acids is enough to alter tree topologies, even in large datasets. Similarly, maximum likelihood analyses of the nucleotide 100 gene, 26 taxon dataset supports hornworts as the sister-group to the remaining land plants (Fig. 1a), but when the data are codon-degenerated the

same analyses result in a monophyletic bryophytes (Fig. 1b). Although these results by themselves do not negate the support for the hornworts as the sister-lineage to the remaining land plants in the nucleotide data, they do suggest that that support is due entirely to the faster-evolving synonymous substitutions that are problematic to model due to increased rates of substitution and the accumulation of composition biases.

The importance of using non-stationary substitution models

In this study we analysed a 100 protein-coding gene and 26 taxon dataset obtained from a larger previously published 620 gene, 103 taxon dataset of nuclear gene sequences. This reduced dataset was generated so that evolutionary models that account for composition heterogeneity could be used but which are computationally intractable on larger datasets. Such a methodology is based on the supposition that modeling the substitution process is an equally important part of the practice of phylogenetics as is taxon sampling. In the era of next-generation sequencing techniques and the ease of obtaining vast amounts of comparative sequence data, it can be argued that taxon sampling is no longer the limiting factor in phylogenetic systematics, but rather our ability to model the complexity of the evolutionary process. Indeed, adequate taxon sampling is not dependent merely on numbers of taxa but rather upon a judicious taxon sampling needed to address the specific relationships the analyses are aimed at resolving (Cox 2014). For instance, if the analyses are aimed at resolving relationships among the three bryophyte groups, then it is more important to sample lineages that represent temporally sparse phylogenetic splits in each bryophyte group, such as the moss genera *Takakia* and *Andreaea*, than it is to sample densely within evolutionary-derived taxa such as the speciose pleurocarpous moss group Hypnanae. Including many such taxa would be superfluous while limiting the complexity of the models that can be used, due to computational constraints. A balance needs to be made between data set size and model complexity, and if analyses with large taxon samples can only apply simplified models that ignore heterogeneity and fit the data poorly, they should be treated with due skepticism.

The criteria used to select taxa and genes for the reduced (100 genes, 26 taxa) data set were aimed at decreasing the effect of biological sources of phylogenetic incongruence such as elevated rates of substitution, by preferring shorter gene trees, and at minimising composition heterogeneity among taxa. Nevertheless, the synonymous to non-synomymous substitution rate of the 100 chosen genes ranged from 6-300X, indicating that our selection procedure had little effect on limiting the influence of the fast-evolving synonymous substitutions on the analyses, compared to the full 620 gene data set. Moreover, the selected data that comprised the reduced data set were

not composition homogeneous even if the amount of heterogeneity was reduced: posterior predictive distribution of the X^2 of composition homogeneity p=0.0 (Fig. S8). Indeed, despite our attempts to reduce possible sources of phylogenetic artifacts, our reduced data set had very similar analytical characteristics as the full 620 gene data set.

Using better-fitting composition heterogeneous models did not alter the inferred topology or the support, compared to homogeneous models, when analysing either the nucleotide or codon-degenerate alignments, although the former supported hornworts as the sister-group to all land plants whereas the latter a monophyletic bryophytes (Fig. S8 vs. Fig. S9 $2\log_e$ BF=9016.7151 and Fig. S11 vs. Fig. S12 $2\log_e$ BF=961.2782, respectively). Among-lineage composition heterogeneity is present in the nucleotide data but its modeling has no influence on the phylogenetic result, indicating there are other processes that have a larger and overwhelming impact on the analyses. In contrast, when analysing the more-slowly evolving amino acid data, using a better-fitting composition heterogeneous model does increase branch support for a monophyletic bryophyte group significantly (PP=0.98, Fig. 2d), compared to the homogeneous model (PP=0.84, Fig. 2C). We speculate that, because amino acids have a greater number of potential identities (n=20) when compared to nucleotides (n=4), there is greater potential for variation in among-lineage composition heterogeneity and therefore modeling composition biases has a greater effect in amino acid data.

Implications of the study for understanding the evolution of land plants

Composition heterogeneity in nuclear land plant molecular data has been shown to affect the inference of phylogenetic relationships in analyses of poorly-fitting homogeneous (stationary) composition models. Indeed, the best-fitting composition models found for the nucleotide data, the codon-degenerate nucleotide data, and the amino acid data, were all heterogeneous, indicating that any analyses of these data under homogeneous composition models is highly questionable. Analyses of the codon-degenerate nucleotide data and the amino acid data using the best-fitting non-stationary composition models resolve the bryophytes as monophyletic group with high branch support. Our results from nuclear protein-coding gene data provide compelling evidence that the three lineages of bryophytes, mosses, liverworts, and hornworts, form a monophyletic group and thereby share a common ancestor to the exclusion of tracheophytes. This hypothesis implies that the first phylogenetic split among land plants was between the bryophytes and tracheophytes, rather than the tracheophytes being derived from bryophyte ancestors, which has been the prevailing theory. These results are congruent with recently published studies of

chloroplast (Nishiyama *et al.* 2004; Cox *et al.* 2014) and nuclear (Puttick et al 2018) protein-coding genes that favour the monophyly of bryophytes over other possible resolutions of the land plant phylogeny (Cox *et al.* 2014; Puttick *et al.* 2018). In addition, the Setaphyta (Puttick *et al.* 2018), the clade consisting of mosses and liverworts, is recovered in all but one analysis. The study of Puttick *et al.* (2018) which also re-analysed the amino acid data of Wickett *et al.* (2014), strongly favoured the monophyly of bryophytes, the clade being highly supported in several analyses including supertree analyses from gene trees and composition-heterogeneous analyses of Dayhoff groups. However, using a reduced low-heterogeneity dataset and a jack-knife approach, the alternative topologies that place hornworts either as the sister-group to the other embryophytes or as the sister-group to the tracheophytes could not be rejected. Here, we focus instead on direct comparisons between analyses of nucleotide, codon-degenerate nucleotide, and amino acid data of the same 100 gene dataset, and between inferences under composition tree-homogeneous and tree-heterogeneous models, showing that when codon degeneracy and non-stationary models are used, inferences from both nucleotide and amino acid data converge on the same topology supporting the monophyly of bryophytes. Indeed, the explanation that incongruence between analyses of nucleotide protein-coding gene data and their amino acid translations is due to fast-evolving (and therefore unreliable) synonymous substitutions was also given for similar incongruences among analyses of land plant chloroplast data; data that were also shown to best support a monophyletic bryophytes (Cox *et al.* 2014). Consequently, the hypothesis that bryophytes are monophyletic is now better supported than alternatives indicating bryophyte paraphyly.

A common origin of bryophytes has profound implications for the way that land plant evolution is understood. For instance, it challenges the fundamental idea that the bryophyte life cycle, in which the gametophyte is the dominant vegetative stage and nurtures an unbranched sporophyte, is ancestral to land plants (Haig 2008). Indeed, although the haplobiontic life-cycles (with dominant gametophytes and zygotic meiosis) of the charophyte algal ancestors of land plants imply that the gametophyte of the land plant ancestor was multicellular, given the monophyly of both bryophytes and tracheophytes, it is possible that the sporophyte of the ancestor of land plants was branched, and maybe even the dominant phase of the life-cycle as in tracheophytes. In such a case, the unbranched sporophyte of the bryophytes would represent a reduction from the more elaborate ancestral sporophyte. Moreover, assuming homology between the retention of the meiotic zygotes in the oogonia of the haploid phase of such charophytes as *Chara ssp.* and the nurturing of the sporophyte by the haploid gametophyte of bryophytes, the ancestor of land plants likely had a sporophyte attached to, or nourished by, the gametophyte. However, if this assumption of homology is incorrect, the most recent common ancestor of land plants may have had independent

gametophytes and sporophytes that were near-isomorphic, or with either phase being dominant, and the dependence of the sporophyte upon the gametophyte may be a a derived character of the bryophyte lineage. Another corollary to the acceptance of bryophyte monophyly over other evolutionary scenarios is that the presence of stomata is likely a synapmorphy of all embryophytes and present in the ancestral sporophyte of all land plants, and subsequently lost in the liverwort lineage. Earlier phylogenetic hypotheses that placed liverworts as the sister-group to all other embryophytes implied that stomata arose in the embryophyte lineage after the divergence of liverworts.

Taxonomy of a monophyletic bryophytes

The clade uniting all three bryophyte lineages should be referred to by its formal name in accordance with taxonomic precedence. The name Bryophyta *sensu lato* has been used informally to refer to all bryophytes (Cronquist *et al.* 1966; Whittaker 1969), but using it as a formal name creates ambiguity with Bryophyta *sensu stricto*, which pertains only to mosses (Goffinet and Buck 2013; Ruggiero *et al.* 2015). The name “Bryobiotina“ has previously been proposed for a subkingdom encompassing all three bryophyte lineages (Campbell 1891). However, assigning the rank of subkingdom to the bryophytes is problematic, as there are several unranked taxa within the kingdom Plantae, such as Streptophyta and Embryophyta, that include the bryophytes. Furthermore, the sister-lineage to all bryophytes, Tracheophyta, is also an unranked taxon. We propose that the previously used division (phylum) name Bryophyta Schimp. (1879) be used for the clade containing mosses, liverworts, and hornworts. This will give taxonomic symmetry to the land plant classification with the first split being between the Tracheophyta and Bryophyta. Schimper originally used the name Bryophyta to describe both the mosses and liverworts (which at the time included the hornworts). More recently, the name Bryophyta Schimp. has been restricted in use to the mosses alone (e.g. Goffinet *et al.* 2009), with the liverworts (Marchantiophyta Stotler & Crand.-Stotl.) and hornworts (Anthoceratophyta Rothm. Stotler & Crand.-Stotl.) recognised as separate divisions. The elevation of the three bryophyte lineages to individual divisions was done presumably to reflect the concept of the paraphyly of bryophytes. If the monophyly of bryophytes is to be recognised it seems now prudent to de-rank the hornworts, liverworts and mosses, to the classes Anthocerotopsida, Marchantiopsida, and Bryopsida respectively, and classify the bryophytes as a whole as Bryophyta.

Acknowledgments

This study was funded by FCT (Portuguese Foundation for Science and Technology) through project grant PTDC/BIA-EVF/1499/2014 to C.J.C. and institutional grant CCMAR/Multi/04326/2013, and by the NERC (Natural Environment Research Council, U.K.) grant NE/N002067/1 to P.D., and H.S. We also wish to thank the Portuguese Infraestrutura Nacional de Computação Distribuída for access to their High Performance Computing infrastructure (INGRID).

Author contributions

C.J.C.,P.G.F., P.C.J.D., and H.S. conceived the study. F.S. and C.J.C. performed analyses. F.S., P.G.F., P.C.J.D., H.S. and C.J.C. wrote the paper.

References

Blanquart S, Lartillot N. 2008. A site- and time-heterogeneous model of amino acid replacement. *Molecular Biology and Evolution* 25: 842–858.

Bulmer M. 1988. Are codon usage patterns in unicellular organisms determined by selection-mutation balance? *Journal of Evolutionary Biology* 1: 15–26.

Campbell DH. 1891. *Elements of Structural and Systematic Botany.* Boston, U.S.A.: Ginn & Company.

Chang Y, Graham SW. 2011. Inferring the higher-order phylogeny of mosses (Bryophyta) and relatives using a large, multigene plastid data set. *American Journal of Botany* 98: 839–849.

Chernomor O, von Haeseler A, Minh BQ. 2016. Terrace aware data structure for phylogenomic inference from supermatrices. *Systematic biology* 65: 997–1008.

Civan P, Foster PG, Embley MT, Seneca A, Cox CJ. 2014. Analyses of charophyte chloroplast genomes help characterize the ancestral chloroplast genome of land plants. *Genome Biology and Evolution* 6: 897–911.

Clarke JT, Warnock RCM, Donoghue PCJ. 2011. Establishing a time-scale for plant evolution. *New Phytologist* 192: 266–301.

572 Cox CJ, Foster PG, Hirt RP, Harris SR, Embley TM. 2008. The archaebacterial origin of
573 eukaryotes. *Proceedings of the National Academy of Sciences* 105: 20356–20361.

574 Cox CJ, Li B, Foster PG, Embley TM, Civáň P. 2014. Conflicting phylogenies for early land plants
575 are caused by composition biases among synonymous substitutions. *Systematic Biology* 63:
576 272–279.

577 Cox CJ. 2018. Land plant molecular phylogenetics: a review with comments on evaluating
578 incongruence among phylogenies. Critical Reviews in Plant Sciences 1-15.
579 DOI:10.1080/07352689.2018.1482443

580 Cronquist A, Takhtajan A, Zimmermann W. (1966). On the higher taxa of Embryobionta. *Taxon* 15:
581 129-134.

582 Dayhoff MO, Schwartz RM, Orcutt BC. 1978. A model of evolutionary change in proteins. In:
583 Dayhoff MO, ed. Atlas of Protein Sequence and Structure, vol. 5. Washington DC, U.S.A.:
584 National Biomedical Research Foundation, 345-352.

585 Foster PG. 2004. Modeling compositional heterogeneity. *Systematic Biology* 53: 485–495.

586 Foster PG, Cox CJ, Embley TM. 2009. The primary divisions of life: A phylogenomic approach
587 employing composition-heterogeneous methods. *Philosophical Transactions of the Royal*
588 *Society B: Biological Sciences* 364: 2197–2207.

589 Gao L, Su YJ, Wang T. 2010. Plastid genome sequencing, comparative genomics, and
590 phylogenomics: Current status and prospects. *Journal of Systematics and Evolution* 48: 77–
591 93.

592 Goffinet B, Buck WR, Shaw AJ. 2009. Morphology, anatomy, and classification of the Bryophyta.
593 In Goffinet B and Shaw AJ, eds. *Bryophyte biology*. Cambridge, Cambridge University Press,
594 55 – 138.

595 Goffinet B, Buck WR. 2013. The evolution of body form in bryophytes. *Annual Plant Reviews* 45:
596 51-89.

597 Gouy M, Gautier C. 1982. Codon usage in bacteria: correlation with gene expressivity. *Nucleic*
598 *acids research* 10. 7055–7074.

599 Gouy M, Guindon S, Gascuel O. 2009. SeaView version 4: a multiplatform graphical user interface
600 for sequence alignment and phylogenetic tree building. *Molecular biology and evolution* 27.
601 221–224.

602 Graham LKE, Wilcox LW. 2000. The origin of alternation of generations in land plants: A focus on
603 matrotrophy and hexose transport. *Philosophical Transactions of the Royal Society B:*
604 *Biological Sciences* 355: 757–767.

605 Haig D. 2008. Homologous versus antithetic alternation of generations and the origin of
606 sporophytes. The Botanical Review 74: 395-418.

607 Huang H, Knowles LL. 2009. What is the danger of the anomaly zone for empirical phylogenetics?
608 *Systematic Biology* 58: 527–536.

609 Inagaki Y, Simpson AG, Dacks JB, Roger AJ. 2004. Phylogenetic artifacts can be caused by leucine,
610 serine, and arginine codon usage heterogeneity: dinoflagellate plastid origins as a case study.
611 *Systematic biology* 53: 582–593.

612 Inagaki Y, Roger AJ. 2006. Phylogenetic estimation under codon models can be biased by codon
613 usage heterogeneity. *Molecular Phylogenetics and Evolution* 40: 428–434.

614 Jeffroy O, Brinkmann H. 2006. Phylogenomics: the beginning of incongruence? *TRENDS in*
615 *Genetics* 22: 225-231.

616 Karol KG. 2001. The Closest Living Relatives of Land Plants. *Science:* 294: 2351–2353.

617 Karol KG, Arumuganathan K, Boore JL, Duffy AM, Everett KDE, Hall JD, Hansen SK, Kuehl JV,
618 Mandoli DF, Mishler BD *et al.* 2010. Complete plastome sequences of *Equisetum arvense* and
619 *Isoetes flaccida*: implications for phylogeny and plastid genome evolution of early land plant
620 lineages. *BMC evolutionary biology* 10: 321.

621 Keane TM, Creevey CJ, Pentony MM, Naughton TJ, Mclnerney JO. 2006. Assessment of methods
622 for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions
623 for choice of matrix are not justified. *BMC evolutionary biology* 6: 29.

624 Kenrick P, Wellman CH, Schneider H, Edgecombe GD. 2012. A timeline for terrestrialization:
625 consequences for the carbon cycle in the Palaeozoic, *Philosophical Transactions of the Royal*
626 *Society B: Biological Sciences* 367: 519–536.

627 Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for
628 phylogenetic reconstruction and molecular dating. *Bioinformatics* 25: 2286–2288.

629 Laurin-Lemay S, Brinkmann H, Philippe H. 2012. Origin of land plants revisited in the light of
630 sequence contamination and missing data. *Current Biology* 22: R593–R594.

631 Lenton TM, Crouch M, Johnson M, Pires N, Dolan L. 2012. First plants cooled the Ordovician.
632 *Nature Geoscience* 5: 86–89.

633 Ligrone R, Duckett JG, Renzaglia KS. 2012. Major transitions in the evolution of early land plants:
634 A bryological perspective. *Annals of Botany* 109: 851–871.

635 Liu Y, Cox CJ, Wang W, Goffinet B. 2014. Mitochondrial phylogenomics of early land plants:
636 Mitigating the effects of saturation, compositional heterogeneity, and codon-usage bias.
637 *Systematic Biology* 63: 862–878.

638 Lockhart PJ, Howe CJ, Bryant DA, Beanland TJ, Larkum AWD. 1992. Substitutional bias
639 confounds inference of cyanelle origins from sequence data. *Journal of molecular evolution*
640 34: 153–162.

641 Magallón S, Hilu KW, Quandt D. 2013. Land plant evolutionary timeline: gene effects are
642 secondary to fossil constraints in relaxed clock estimation of age and substitution rates.
643 *American Journal of Botany* 100: 556-573.

644 McCourt RM, Delwiche CF, Karol KG. 2004. Charophyte algae and land plant origins. *Trends in*
645 *Ecology and Evolution* 19: 661–666.

646 Mooers AØ, Holmes EC. 2000. The evolution of base composition and phylogenetic inference.
647 *Trends in Ecology & Evolution* 15: 365–369.

648 Morris JL, Puttick MN, Clark JW, Edwards D, Kenrick P, Pressel S, Wellman CH, Yang Z,
649 Schneider H, Donoghue PCJ. 2018. The timescale of early land plant evolution. *Proceedings*
650 *of the National Academy of Sciences* 115: E2274-E2283.

651 Newton MA, Raftery AE. 1994. Approximate Bayesian inference with the weighted likelihood
652 bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)* 56: 3–48.

653 Nguyen LT, Schmidt HS, von Haeseler A, Minh BQ. 2015. IQ-TREE: A fast and effective stochastic
654 algorithm for estimating maximum-likelihood phylogenies, *Molecular Biology and Evolution*
655 32: 268–274.

656 Niklas KJ, Kutschera U. 2010. The evolution of the land plant life cycle. *New Phytologist* 185: 27-
657 41.

658 Nishiyama T, Wolf PG, Kugita M, Sinclair RB, Sugita M, Sugiura C, Wakasugi T, Yamada K,
659 Yoshinaga K, Yamaguchi K *et al.* 2004. Chloroplast phylogeny indicates that bryophytes are
660 monophyletic. *Molecular Biology and Evolution* 21: 1813–1819.

661 Nishiyama T, Kato M. 1999. Molecular phylogenetic analysis among bryophytes and tracheophytes
662 based on combined data of plasmid coded genes and the 18S rRNA gene. *Molecular Biology*
663 *and Evolution* 16: 1027–1036.

664 Nishiyama T, Sakayama H, de Vries J, Buschmann H, Saint-Marcoux D, Ullrich KK, Haas FB,
665 Vanderstraeten L, Becker D, Lang D *et al.* 2018. The Chara Genome: secondary complexity
666 and implications for plant terrestrialization. *Cell* 174: 448-464.

667 Nylander JA, Ronquist F, Huelsenbeck JP, Nieves-Aldrey J. 2004. Bayesian phylogenetic analysis
668 of combined data. *Systematic biology* 53: 47–67.

669 Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Wörheide G, Baurain D. 2011.
670 Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS*
671 *biology* 9: e1000602.

672 Philippe H, Laurent J. 1998. How good are deep phylogenetic trees? *Current Opinion in Genetics*
673 *and Development* 8: 616–623.

674 Plotkin JB, Kudla G. 2011. Synonymous but not the same: the causes and consequences of codon
675 bias. *Nature Reviews Genetics* 12: 32–42.

676 Puttick MN, Morris JL, Williams TA, Cox CJ, Edwards D, Kenrick P, Pressel S, Wellman CH,
677 Schneider H, Pisani D, Donoghue, PC. (2018). The interrelationships of land plants and the
678 nature of the ancestral embryophyte. *Current Biology* 28: 733-745.

679 Qiu Y-L, Li L, Wang B, Chen Z, Knoop V, Groth-Malonek M, Dombrovska O, Lee J, Kent L, Rest
680 J, *et al.* 2006. The deepest divergences in land plants inferred from phylogenomic evidence.
681 *Proceedings of the National Academy of Sciences* 103: 15511–15516.

682 Regier JC, Shultz JW, Zwick A, Hussey A, Ball B, Wetzer R, Martin JW, Cunningham CW. 2010.
683 Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding
684 sequences 11: 1–6.

685 Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard
686 MA, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and
687 model choice across a large model space. *Systematic biology* 61: 539–542.

688 Rota-Stabelli O, Lartillot N, Philippe H, Pisani D. 2012. Serine codon-usage bias in deep
689 phylogenomics: pancrustacean relationships as a case study. *Systematic biology* 62: 121–133.

690 Ruggiero MA, Gordon DP, Orrell TM, Bailly N, Bourgoin T, Brusca RC, Cavalier-Smith T, Guiry
691 MD, Kirk PM. 2015. A higher level classification of all living organisms. PloS one, 10(4),
692 e0119248.

693 Schimper WP. 1879. Bryophyta. In: von Zittel KA, eds. Handbuch der palaeontologie Vol. 2. R.
694 München & Leipzig: R. Oldenbourg, 152 pp.

695 Sharp PM, Stenico M, Peden JF, Lloyd AT. 1993. Codon usage: mutational bias, translational
696 selection, or both? *Biochemical Society transactions* 21: 835–41.

697 Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
698 phylogenies. *Bioinformatics* 30: 1312-1313.

699 Stamatakis A, Aberer AJ. 2013. Novel parallelization schemes for large-scale likelihood-based
700 phylogenetic inference. In: Proceedings of the 2013 IEEE 27th International Symposium on
701 Parallel and Distributed Processing, IEEE Computer Society, 1195–1204.

Stebbins GL, Hill GJC. 1980. Did multicellular plants invade the land? *The American Naturalist* 115: 342–353.

Stenøien HK. 2005. Adaptive basis of codon usage in the haploid moss *Physcomitrella patens.* *Heredity* 94: 87.

Susko E, Roger AJ. 2007. On reduced amino acid alphabets for phylogenetic inference. *Molecular Biology and Evolution* 24: 2139–2150.

Timme RE, Bachvaroff TR, Delwiche CF. 2012. Broad phylogenomic sampling and the sister lineage of land plants. *PLoS ONE* 7: e29696.

Whittaker RH. 1969. New concepts of kingdoms of organisms. Science, 163(3863), 150-160.

Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter C, Matasci N, Ayyampalayam S, Barker MS, Burleigh JG, Gitzendanner MA *et al.*. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences* 111: E4859–E4868.

Wodniok S, Brinkmann H, Glöckner G, Heidel AJ, Philippe H, Melkonian M, Becker B. 2011. Origin of land plants: Do conjugating green algae hold the key? *BMC Evolutionary Biology* 11: 104.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution* 24: 1586–1591.

Yang Z, Nielsen R. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *Journal of molecular evolution* 46: 409–418.

Zhou M, Li, X. 2009. Analysis of synonymous codon usage patterns in different plant mitochondrial genomes. *Molecular biology reports* 36: 2039-2046.

Figure legends

Figure 1: Majority-rule consensus trees inferred from the 100 gene, 26 taxon concatenated nucleotide data set. a) Majority-rule consensus tree of maximum likelihood bootstrap analyses (300 replicates) under the GTR+G4+F_{est} model, b) the corresponding analysis of codon-degenerated data under the same model. Taxa are indicated as follows: hornworts – orange, liverworts – cyan blue, mosses – light green, tracheophytes – violet.

Figure 2: Majority-rule consensus trees inferred from the 100 gene, 26 taxon concatenated amino acid data. a) maximum likelihood bootstrap with 300 replicates under the model LG+G_4+F_{est}, b) maximum likelihood bootstrap analysis with 300 replicates of the Dayhoff-recoded data with under the model GTR+G_4+F_{est}, c) Bayesian MCMC of the amino acid data with a composition homogeneous model LG+G_4+F_{CV1}, marginal likelihood -L_h=441823.4926, d) Bayesian MCMC of the amino acid data with a heterogeneous NDCH composition model LG+G_4+F_{CV5}, marginal likelihood -L_h=441066.4929. Taxa are indicated as follows: hornworts – orange, liverworts – cyan blue, mosses – light green, tracheophytes – violet.

Supporting Information Legends

Notes S1: Calculation of non-synonymous/synonymous substitution rates for 85 genes from the 620 gene data set.

Figure S1: Reanalyses of the 620 genes, 103 taxa nucleotide dataset, RAxML full bootstrap, GTRGAMMA, 200 replicates.

Figure S2: Reanalyses of the 620 genes, 103 taxa nucleotide dataset without 3rd-codon positions, RAxML full bootstrap, GTRGAMMAX, 200 replicates.

Figure S3: Reanalyses of the 620 genes, 103 taxa nucleotide dataset, codon-degenerate without 3rd-codon positions, RAxML full bootstrap, GTRGAMMAX, 200 replicates.

Figure S4: Reanalyses of the 620 genes, 103 taxa nucleotide dataset, codon-degenerate without 3rd-codon positions, RAxML full bootstrap, GTRCATX, 200 replicates.

Figure S5: Reanalyses of the 620 genes, 103 taxa amino acid dataset, Partitioned RAxML full bootstrap, PROTCAT(X), 100 replicates.

Figure S6: Reanalyses of the 620 genes, 103 taxa amino acid dataset, Partitioned RAxML full bootstrap, PROTGAMMA(X), 100 replicates.

Notes S2: Calculation of non-synonymous/synonymous substitution rates for 35 genes from the 100 gene data set.

Table S1: The list of 100 nuclear genes showing the sequence length, number of taxa and the number of composition vectors that fits the data for both nucleotide (nt) and amino acid (aa) alignments.

Figure S7: Analyses of the 100 genes, 26 taxa nucleotide dataset, Partitioned IQTREE ML bootstrap (greedy) analysis, with 100 replicates.

Figure S8: Analyses of the 100 genes, 26 taxa nucleotide dataset, Bayesian P4 MCMC, GTR+Gamma, homogeneous composition (CV1).

Figure S9: Analyses of the 100 genes, 26 taxa nucleotide dataset, Bayesian P4 MCMC, GTR+Gamma, heterogeneous composition (CV2).

Figure S10: Analyses of the 100 genes, 26 taxa nucleotide dataset, PhyloBayes MCMC, CAT-GTR+Gamma.

Figure S11: Analyses of the 100 genes, 26 taxa codon-degenerate nucleotide dataset, Bayesian P4 MCMC, GTR+Gamma, homogeneous composition (CV1).

Figure S12: Analyses of the 100 genes, 26 taxa codon-degenerate nucleotide dataset, Bayesian P4 MCMC, GTR+Gamma, heterogeneous composition (CV2).

Figure S13: Analyses of the 100 genes, 26 taxa nucleotide dataset, Codon analysis, IQTREE ML bootstrap, GY2K+F3X4+G, 100 replicates.

Figure S14: Analyses of the 100 genes, 26 taxa nucleotide dataset, Codon analysis, IQTREE ML bootstrap, MG2K+F3X4+G, 100 replicates.

Figure S15: Analyses of the 100 genes, 26 taxa Dayhoff amino acid group dataset, Bayesian P4 MCMC, GTR+Gamma, homogeneous composition (CV1).

Figure S16: Analyses of the 100 genes, 26 taxa Dayhoff amino acid group dataset, Bayesian P4 MCMC, GTR+Gamma, heterogeneous composition (CV2).

Figure S17: Analyses of the 100 genes, 26 taxa amino acid dataset, PhyloBayes MCMC, CAT-LG+Gamma.